Journal of Research on Educational Effectiveness, 8: 149–168, 2015

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2014.906009

Routledge
Taylor & Francis Group

INTERVENTION, EVALUATION, AND POLICY STUDIES

Social Identity and Achievement Gaps: Evidence From an Affirmation Intervention

Thomas S. Dee
Stanford University, Stanford, California, USA, and NBER, Cambridge, Massachusetts, USA

Abstract: One provocative explanation for the continued persistence of minority achievement gaps involves the performance-dampening anxiety thought to be experienced by minority students in highly evaluative settings (i.e., "stereotype threat"). Recent field-experimental studies suggest that modest, low-cost "buffering" interventions informed by this phenomenon may be highly effective at reducing minority achievement gaps. This field-experimental study evaluates such an intervention in which students complete a self-directed "self affirmation" exercise that encourages them to identify and reflect upon their core personal values. This within-classroom randomized trial was conducted among 2,500 7th and 8th graders from six Philadelphia-area middle schools during the 2008–09 and 2009–10 academic years. Although this study failed to replicate the earlier findings indicating that the affirmation generated large increases in the academic performance of minority students, this treatment did lead to statistically significant improvements in the performance of the minority students in more supportive classroom environments. However, the treatment contrast also *reduced* the performance of female students in those settings.

Keywords: Stereotype threat, achievement gaps, field intervention

The large and persistent underperformance of African American and Hispanic students relative to their White peers is a centrally relevant policy concern, both because of its implications for long-run inequality and because of the loss of growth-enhancing human capital. The effects that broad institutions, policies, and practices (e.g., schools, educational resources and incentives, socioeconomic priors and culture) have on these gaps are, justifiably, topics of sustained interest among researchers and policymakers. However, some of the most provocative recent conjectures about determinants of achievement gaps focus on psychological processes related to students' social identity in academic settings. In particular, a recent literature originating in the field of social psychology suggests that a performance-dampening anxiety that can be experienced by minorities in highly evaluative settings (i.e., "stereotype threat") contributes to achievement gaps (Steele & Aronson, 1995). An extensive body of lab-based studies is consistent with this phenomenon, showing

Address correspondence to Thomas S. Dee, Center for Education Policy Analysis, Stanford University, 520 Galvez Mall, CERAS Building, 5th Floor, Stanford, CA 94305-3001, USA. E-mail: tdee@stanford.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree.

that subtle priming of stereotyped social identities can lead to large gaps in measured test performance (Aronson & Dee, 2011; Aronson & McGlone, 2007; Schmader, Johns, & Forbes, 2008).

A particularly provocative and recent extension of this literature has focused on designing and evaluating *field* interventions that seek to "buffer" minority students from this phenomenon and improve their academic performance. These early field-experimental studies have generated an exciting pattern of results suggesting that quite modest, low-cost, and seemingly scalable interventions can lead to surprisingly large gains in the academic performance of minority students (e.g., Arbuthnot, 2009; Cohen, Garcia, Apfel, & Master, 2006; Good, Aronson, & Harder, 2008; Good, Aronson, & Inzlicht, 2003; Keller & Dauenheimer, 2003; Kellow & Jones, 2008). For example, the recent within-classroom random-assignment study by Cohen et al. (2006) found that having seventh-grade students complete a self-directed, 15-min "affirmation" exercise in their classroom improved the final grade of African American students in their "treated" subject by an amount equal to 40% of the Black–White achievement gap (i.e., an effect size as large as 0.34).

This study presents a large-scale field-experimental evaluation of this affirmation exercise. This experiment was conducted among more than 2,500 seventh and eighth graders at six Philadelphia-area middle schools over 2 years. This study constitutes the first independent replication of the provocative results reported by Cohen et al. (2006) in a similar general-education setting. The experimental materials and protocols used in this study closely parallel those of the seminal study. Students within the participating classrooms were randomly assigned to complete either the treatment or control versions of the affirmation exercise. The central research question in this study is whether student completion of this value affirmation influenced their grade in that classroom. As in the original study, a question of particular interest is whether these treatment effects differ by ethnicity or sex. The larger number of classrooms participating in this replication also makes it possible to explore classroom moderators of this treatment. This study also provides evidence on whether the treatment influenced secondary outcomes such as test scores, grades in other subjects, absences, and so on. These and other distinctive features of this replication are discussed next.

SOCIAL IDENTITY AND ACHIEVEMENT GAPS—THEORY AND EVIDENCE

The broad policy interest in minority achievement gaps is motivated in no small part by the evidence that these persistent achievement differences explain a substantial portion of the longer run inequality in educational attainment and labor-market success (e.g., Neal & Johnson, 1996; O'Neill, 1990). The African American–White and Hispanic–White gaps in reading and math achievement narrowed considerably between the early 1970s and the late 1980s. However, that dramatic convergence has largely stalled over the last 20 years. The current achievement gaps remain quite large. For example, data from the 2008 National Assessment of Educational Progress Long-Term Trend study indicate that 13-year-old African American and Hispanic students underperform in math relative to White students by amounts equal to 74% and 66% of a standard deviation, respectively (Rampey, Dion, & Donahue, 2009).

Explanations for the earlier convergence in achievement gaps have often focused on the effects of relative changes in family background and in the quality of schools attended by minority and nonminority students. For example, several recent studies find that

court-ordered school desegregation plans improved the educational attainment and subsequent earnings of African American students (e.g., Guryan, 2004; Johnson, 2011; Reber, 2010), implying that they increased school quality. However, decompositions of the convergence in Black–White achievement gaps (Cook & Evans, 2000) suggest instead that they were largely attributable to a narrowing of performance among African American and White students with similar backgrounds and attending the same schools. This sort of *within-school* convergence in African American–White achievement gaps could reflect changes in schools' tracking and enrollment practices as well as changes in teacher expectations across minority and nonminority students.

Other provocative explanations for the relative academic performance of minority students (and, perhaps, how they have changed over time) have focused on mechanisms related to social identity. For example, one prominent hypothesis is that the academic performance of minority students suffers because of the negative peer stigma associated with "acting White" (Fordham & Ogbu, 1986). However, at least for African American students, the empirical evidence in support of this hypothesis (e.g., whether improved academic performance lowers the popularity of minority students) is limited to high-performing students (Cook & Ludwig, 1997; Fryer & Torelli, 2010).

Another widely discussed explanation for the size and persistence of minority achievement gaps is the social-psychological concept of "stereotype threat" (Steele & Aronson, 1995). Stereotype threat (ST) refers to the claim that, in highly evaluative settings (such as classrooms), individuals can experience anxiety based on the concern that others will view them through the lens of a negative stereotype. This anxiety can effectively become self-fulfilling by impeding academic performance. In an economic model of student effort, ST can be modeled as a negative shock to the production function that maps student effort into valued skills (Dee, 2014). When effort is a complement to native ability in the production of skills, a negative ST shock unambiguously reduces student effort and, by implication, academic outcomes. However, in situations where students see increased effort as a substitute for ability, a negative ability shock due to ST would increase student effort and, perhaps even, student performance (Dee, 2014). This may explain the small number of anomalous findings that female students exerted more effort when confronted with an offensive cartoon that deprecated the ability of females to do math (i.e., they may view the "priming" as absurd and offensive and respond with increased effort). Regardless, an extensive body of lab-experimental evidence (i.e., more than 300 experiments) suggests the empirical relevance of ST (see Aronson & Dee, 2011; Aronson & McGlone, 2007; Schmader et al., 2008, for reviews). Specifically, multiple studies have demonstrated, for different populations, social identities, and tasks, that making subjects aware of a negative (or positive) social identity (i.e., "priming") can correspondingly influence their subsequent academic performance as well as its mediating measures.

A more recent and provocative development in this literature involves a small number of studies that—using experimental methods in field settings—evaluate the effects on student performance of interventions that may reduce the effects of stereotype threat. For example, several studies (e.g., Arbuthnot, 2009, Good et al., 2008; Keller & Dauenheimer 2003; Kellow & Jones, 2008) find that random assignment to test presentations that emphasize the neutrality of an assessment lower measured achievement gaps. The effect sizes associated with these simple neutrality messages are quite large. For example, Keller and Dauenheimer (2003, Table 3) reported a 0.64 *SD* increase in the test performance of females assigned to a "no threat" condition. Good et al. (2003) evaluated two other buffering interventions among seventh graders in a rural Texas school district serving largely low-income and Latino students. One treatment arm focused on tutoring, which emphasized that the focal

point of stereotypes—intelligence—is a malleable not fixed trait. In a second treatment arm, students were encouraged to attribute academic difficulty to an external factor (i.e., adjusting to a school transition), an approach thought to divert students from the cognitively impairing anxiety associated with ST. Both treatment arms significantly increased reading achievement (effect sizes of 0.5 to 0.7) and the math performance of female students (effect size of 1.0).

The study by Cohen et al. (2006) was a field-experimental evaluation of a third type of ST buffer: a 15-min in-class writing exercise designed to protect students from stereotype threat through the affirmation of core personal values and self-integrity. The population in this study consisted of two cohorts of seventh graders ($n = 282$) in the classrooms of three teachers in a single suburban middle school that served middle-income and lower middle-income families. Slightly less than half of the students participating in this study were African American. This study concluded that within-classroom random assignment to the treatment condition had no statistically significant effect on the subsequent grade point average (GPA) of White students but increased the GPA of the African American students by a statistically significant 0.26 (Experiment 1) to 0.34 (Experiment 2) points.[1] Of interest, a follow-up study, which also introduced a third cohort of students ($n = 134$), found that these African American–specific achievement gains persisted into a 2nd year (Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzutoski, 2009). More recent field-experimental studies have examined the effects of a similar affirmation in a variety of settings and, generally, though not always, found positive effects (Bowen, Wegmann, & Webber, 2013; Cook, Purdie-Vaughns, Garcia, & Cohen, 2012; Hanselman, Bruch, Gamoran, & Borman, 2014, Harackiewicz et al., 2014; Lauer et al., 2013; Miyake et al., 2010; Sherman et al., 2013; Woolf, McManus, Gill, & Dacre, 2009).

As noted in the introduction, this study presents a new and independent evaluation of the effects of this writing affirmation, which was conducted in six Philadelphia-area middle schools during the 2008–09 and 2009–10 academic years. The details of this field experiment are described next. However, by way of comparison with the study by Cohen et al. (2006), a few comparative features are worth underscoring. First, the procedures and materials used in this experiment exactly replicated those from the original study. Second, this field experiment was conducted on a larger scale: Roughly 2,500 students in 139 classrooms across six schools who were taught by 22 teachers. Third, some participating schools resembled the single school in the original study (i.e., in terms of the percentage of students who were African American and general socioeconomic traits). However, several participating schools also served sizable populations of Hispanic students. Fourth, in contrast to the original study, virtually all of the participating schools utilized a "passive consent" model. This meant that an unusually large share of eligible students (i.e., nearly 90%) actually participated in this study. In the study by Cohen et al. (2006), only 64% of students returned their consent forms and only 80% of those students provided consent. This implies that only about half of the study-eligible students actually participated. Fifth, in addition to the primary outcome measure (i.e., GPA in the "treated" academic course), this study also includes multiple other outcome measures: posttreatment spring reading and math assessment scores, grades in nontreated academic subjects, attendance, tardiness, and disciplinary data. Finally, the large number of classrooms in this study makes it possible to explore the classroom moderators of treatment efficacy.

[1]Experiment 2 slightly simplified the affirmation, which is described in more detail below. The standard deviation in this course-grade measure was roughly 1.0, so the effect sizes were 0.26 to 0.34.

Table 1. Sample sizes and consenters by school and academic year

Academic Year	Grades	School	District	Baseline Sample Size	Students With Consent	Assigned to Treatment	Assigned to Control
2008–2009	7, 8	A	1	458	431	224	207
2009–2010	7	A	1	209	202	101	101
2008–2009	7, 8	B	1	225	217	112	105
2009–2010	7	B	1	72	72	35	37
2008–2009	7, 8	C	1	237	226	116	110
2009–2010	7	C	1	114	110	60	50
2008–2009	7, 8	D	2	608	306	152	154
2009–2010	7	D	2	351	331	164	167
2009–2010	7, 8	E	3	660	650	324	326
2009–2010	7, 8	F	4	121	120	58	62
Sample size				3,055	2,665	1346	1,319

Note. Passive consent was used in all schools except for school D in Academic Year 2008-09. Among the 2,665 consenting students, 2,564 were White, Black, or Hispanic.

AN AFFIRMATION EXPERIMENT

Settings and Participants

The settings in which the intervention was conducted were the classrooms of a single core academic subject not commonly associated with gender stereotypes. As in the study by Cohen et al. (2006), the exact subject is not provided here to ensure the confidentiality of the participants. However, the intervention was fielded in classrooms during the same academic subject used in the study by Cohen et al. (2006). During the 1st study year (i.e., AY 2008–09), the seventh and eighth graders in four public, Philadelphia-area middle schools (i.e., Schools A through D in Table 1) participated in the study. In the 2nd study year (AY 2009–10), the seventh graders in the original four schools (i.e., A through D) and the seventh and eighth graders in two additional Philadelphia-area middle schools (E and F) participated in the study. Only the seventh graders in Schools A, B, C, and D participated during the 2nd study year, because most of the 2009–10 eighth graders participated during the prior year. For each school and study year, Table 1 provides an overview of the number of students observed at baseline as well as the number of students for whom consent was acquired and who were then blocked by ethnicity and randomly assigned to the treatment or the control condition within classsrooms. In most cases, at least 94% of the students observed at baseline actually participated in the study. This high participation rate reflects the widespread use of a "passive consent" (i.e., opt out) procedure. Specifically, parents and guardians were informed of the study in general terms and given the opportunity to withdraw their child from participation. The only exception to this approach was School D in the 1st study year. There, the participating students were those for whom parents provided active consent. As in Cohen et al. (2006), the participation rate under this opt-in procedure was only 50%. Overall, 87% of study-eligible students (i.e., 2,665 out of 3,055) provided consent.

Of the students who were randomly assigned at baseline, all but 101 were White, African American, or Hispanic ($n = 2{,}564$). I limit the results presented here to this group

Table 2. Baseline traits of participating students by school

School	*n*	% White	% Black	% Hispanic	*n*	Reading Score at Baseline
A	613	0.334	0.475	0.191	568	–0.262
B	285	0.168	0.653	0.179	264	–0.444
C	328	0.241	0.570	0.189	305	–0.401
D	593	0.531	0.437	0.032	565	0.410
E	626	0.593	0.045	0.363	580	0.158
F	119	0.479	0.361	0.160	91	0.020

Note. This sample is based on White, Black, and Hispanic students for whom consent was obtained ($n = 2{,}564$, Table 1). The standardized baseline reading score is from the state's assessment in the prior spring.

of students.[2] Using baseline data from these students pooled over both study years, Table 2 provides information on how these schools differed. All of the schools except School E served a sizable share of African American students (i.e., 36–65%) and Hispanic students (i.e., 16–19%). School E served very few African American students (i.e., 4.5%) but a sizable number of Hispanic students (i.e., 36%). These schools exhibited more heterogeneity in terms of baseline achievement on Pennsylvania's state assessments. Specifically, Schools A, B, and C had comparatively low levels of reading achievement (e.g., 0.26–0.44 *SD*s below the mean defined for this sample), whereas School D had a relatively high level of baseline reading achievement. It should be noted that, relative to the schools statewide, this heterogeneity implies that these schools were generally at or below mean performance levels.

Experimental Procedures

Just prior to the beginning of each academic year, the participating teachers were given an overview of the basic logistics of the study that included a script for introducing the 10–15 writing exercise to students and suggested responses to questions students might ask.[3] The teachers were aware of the basic structure of the assignment but were blind both to the overall intent of the study and to the specific treatment status of individual students. Students received the writing assignments in closed envelopes or folders, and the assignments were designed to be self-directed. The treatment and control assignments, which are described in more detail next, also had virtually identical three-page layouts, implying that the status of individual students could not be easily observed. Writing assignments were given to students twice during each study year (i.e., as in Experiment 2 in Cohen et al., 2006); once in the first few weeks of the academic year and a second time roughly 6 to 8 weeks later. Because students were sometimes absent when the intervention was administered, teachers were instructed to have such students complete the assignment on their return.

The experimental procedure randomly assigned participating students to the treatment and control conditions within participating classrooms. However, to reduce the chance that

[2] As described next, the randomization procedure blocked on race and ethnicity within classrooms so this sample construction does not influence random assignment to condition.

[3] This script was also used in the study by Cohen et al. (2006) and was generously provided by Geoffrey Cohen.

simple random assignment did not balance outcome-relevant traits across the treatment and control conditions, the random-assignment procedure also "blocked" on race, gender, and baseline achievement. Specifically, within participating classrooms, I paired each student with another student of the same race-ethnicity and gender and similar baseline reading scores. I then randomly assigned students within these matched pairs. Students who could not be matched were randomized as singletons. As a check on this procedure, I examined auxiliary regressions where treatment status is the dependent variable and the baseline traits are the independent variables. Random assignment, particularly in light of the fact that I blocked on these traits, should ensure that these traits are balanced by treatment status. The results indicated, across different specifications, that race, gender, and baseline achievement do not have statistically significant relationships with treatment status. In fact, in each of these regressions, the hypothesis that these observables are jointly significant could not be rejected. These results are consistent with the maintained assumption that the random-assignment procedures appear to have performed quite well and suggest that the treatment contrast has a strong causal warrant.

The Treatment Contrast

The writing assignments were first provided to all participating students as closely as possible to the beginning of the academic year (and, if possible, just prior to a quiz or assessment), a time at which "evaluative stress" is thought to be exceptionally high. In both the treatment and the control conditions, each student's three-page packet began by stating that they would be answering questions about "your ideas, your beliefs, and your life" and that it was important to know that there were no right or wrong answers to these questions. The first page of all worksheets listed 11 values (e.g., "Creativity," "Independence," "Living in the Moment," "Relationships with Friends or Family").[4] Students assigned to the treatment condition were asked to circle the two or three values "most important" to them. In the control condition, students were asked to identify the two or three values "least important" to them. On the second page, treatment students were asked to think about the values they had identified and to write "a few sentences" about why they felt they were important, focusing on their thoughts and feelings and not worrying about grammar or spelling. In the control condition, students were asked to write about why *someone else* might find those values important. In the treatment, the third and final page sought to reinforce the affirmation in two ways. First, students were asked to list the two top reasons the chosen values were important to them. Then they were asked to indicate their level of agreement with four positive statements about their values (e.g., "These values have influenced my life" and "I care about these values"). In the control version of the third page, students were asked to identify the top two reasons these values would matter to someone else and to indicate a level of agreement with four statements about how others viewed these values (e.g., "These values have influenced some people" and "Some people care about these values.").

An important question is the extent to which students actually completed these assignments and, by implication, experienced the treatment contrast intended by the study design. For example, some absentee students may have never completed the intervention. Similarly, some students who were asked to complete the exercise may have simply left

[4]I followed the materials used in Experiment 2, which generated larger treatment effects. The first intervention excluded the values "Being Smart or Getting Good Grades," but these values were included in the intervention fielded later in the year.

the assignment blank or declined to complete it fully. To have objective, empirical answers to these important questions, I examined the response rates of the participating students. Fortunately, these data indicate that a high percentage of the students who were randomly assigned at baseline did meaningfully engage in the intervention assignments. Specifically, roughly 94% of the analytical sample had complete responses to the first-page questions on values and completed the subsequent short writing assignment.[5] For the last two sets of questions, the response rates fell to only around 85%. For the second administration of the worksheet, the response rates were only somewhat lower (i.e., 80–88%). Overall, these results confirm that the uptake of the treatment and the existence of a treatment contrast were both strong.

Another salient dimension to the treatment contrast is the extent to which students in the treatment condition actually engaged the critical essay portion of the assignment. Counts of the words completed indicate that students in the treatment condition wrote an average of 66 words. The exact nature of the comments provides some anecdotal sense of how an affirmation, conducted at a possibly stressful time, might reduce anxiety and improve performance. Specifically, examples of student responses on the essays, without corrections for spelling or grammar, include the following:

> "Being good at art is something good to have because you can express your feelings through art. . . .If you have a good relationship with your friends and family they will help you through tough times."
>
> "My third value is to be smart and get good grades. If I get good grades, I can go to a good college, and have a good job when I become an adult."
>
> "These values are important to me because they make up part of me . . . I do everything in my power for these values to play a role in my life"
>
> "These values are the touchstone of my character. It makes me who I am, and I wouldn't want it any other way."
>
> "I work really hard but if I don't get a good grade I just keep trying and don't give up."
>
> "Being smart and getting good grades is important because most people don't even make it to the eighth grade and I don't wanna be one of those people."
>
> "My relationships with family and friends has let me be who I want to be."
>
> "I like to do my own thing and being able to pick my choices in life. . .nobody can tell me who I can and cannot be because its my life."

Outcome Measures

The main outcome measure in this study is the student's final grade in the class in which the intervention occurred. The grades are on a 0-to-100 scale, but some schools reported grades slightly in excess of 100 reflecting extra credit (Table 3). The standard deviation for the

[5]The limited noncompliance with the experimental assignment occurred because students switched classrooms after their baseline assignment or because they were absent and did not complete the intervention worksheet on their return.

Table 3. Summary statistics for outcome measures

Outcome Measures	M	Minimum	Maximum	Sample Size
Grade in treated subject	81.9 (10.8)	37	101	2,348
Grade in Other Subject A	81.3 (10.2)	52	100	2,329
Grade in Other Subject B	81.5 (9.8)	51	101	2,305
Grade in Other Subject C	80.8 (10.5)	37	101.8	2,334
Post-reading assessment	0 (1.0)	–2.5	3.9	2,381
Post-math assessment	0 (1.0)	–2.6	4.2	2,376
Absences	9.2 (8.3)	0	95	1,799
Tardies	6.7 (10.9)	0	111	1,799
Disciplinary infractions	0.9 (2.0)	0	18	1,718
Stereotype-themed words	3.4 (1.2)	0	7	2,077

Note. The analytical sample consists of 2,564 observations. Data on absences, tardies, and disciplinary infractions were not available for Schools D and F.

main outcome measure is 10.8. Additional outcome measures include final grades in three other nontreated academic subjects and each student's standardized, posttreatment scores on the Pennsylvania System of School Assessment reading and math tests. All participating schools with the exception of Schools D and F provided count data on student absences, tardiness, and disciplinary infractions.

The final measure listed in Table 3 reflects a conjectured mediator of the affirmation intervention: the number of "stereotype-themed" words students entered as part of a single-page worksheet completed in the spring of each study year (i.e., roughly six months after the first intervention). This validated worksheet, which was also utilized in the study by Cohen et al. (2006), consisted of 34 word fragments, seven of which could be completed in stereotype-themed ways. An example is "–ACE," which could be completed as "RACE," "FACE," and so on. The variable based on student responses is the number (i.e., from 0 to 7) of candidate words that were completed in a stereotype-relevant manner. The motivation for this measure is that it is thought to capture cognitive activation of racial stereotypes. Cohen et al. (2006) found that African American students in the treatment condition generated fewer stereotype-themed words than African American students in the control condition (i.e., a statistically significant reduction of 0.46 words).

RESULTS

Main Impact Estimates

The general specification used here to examine treatment effects takes the following form:

$$Y_{ic} = \alpha + \beta T_{ic} + \gamma X_{ic} + \mu_c + \varepsilon_{ic}, \tag{1}$$

where T_{ic} is an indicator for random assignment to the treatment condition (i.e., intent to treat), X_{ic} represents individual-level observables, and μ_{c} represents a classroom fixed effect. Standard errors are adjusted to allow for within-classroom dependence in the error term (i.e., "clustering") using the robust variance procedure introduced by Liang and Zeger (1986). Standard errors based on "block" bootstrapping at the classroom level generate

similar results. To explore the robustness of the results based on Equation 1, some specifications rely on alternative controls. Also, as in the study by Cohen et al. (2006), the treatment indicator is also interacted with the race and gender indicators in some specifications.

Table 4 reports, for seven different specifications, the estimated effects of random assignment to the treatment condition (and other observed traits) on the main outcome measure: final grades in the treated subject. These estimates consistently indicate that there are substantive achievement gaps in this measure of student performance, not unlike those seen in test-score data. More specifically, in models that condition on classroom fixed effects (i.e., Models 4–7), the grades of African American and Hispanic students are 7 to 9 points lower, respectively. Given a standard deviation of 10.8, these gaps imply effect sizes of 0.66 and 0.85, respectively. Similarly, girls outperform boys by roughly 4 points (i.e., an effect size of 0.35).

Random assignment to the treatment condition had quite small and no statistically significant effects. For example, Model 4 indicates an intent-to-treat impact of 0.041, a fraction of a grade point. This estimate is also quite precise statistically. The 95% upper confidence limit is 0.77, implying an effect size of no more than 0.07. Allowing the treatment effects to differ by race, ethnicity and gender (as in the original study) leads to similar results. That is, across multiple specifications that condition on school fixed effects, classroom fixed effects, Race × Gender interactions, and baseline reading scores, the ITT estimates were small and not statistically significant for race-ethnicity and gender subgroups. It should be noted that no multiple-comparison corrections were applied to these confirmatory inferences (Schochet, 2008). If they had been, the degree to which these results are consistent with the null hypothesis of no effect would only be increased. A recent literature has also stressed the importance of estimating treatment effects for different quantiles of the outcome distribution rather than just mean impacts (e.g., Bitler, Gelbach, & Hoynes, 2006). I have estimated treatment effects for centiles of the outcome distribution and, for virtually every Subgroup × Centile estimate, it leads to null results consistent with the mean impacts reported here.

It is particularly informative and interesting to note the precision of these impact estimates and how they compare to the results in the original study by Cohen et al. (2006). More specifically, that study reported that the same treatment (Experiment 2) improved the grades of African American students by effect sizes of 0.34. In this study, the effect size for African American students (Model 5, Table 4) is not statistically significant ($d = 0.02$, 0.228/10.8). The 95% upper confidence limit on this African American–specific treatment effect is 0.14 (i.e., 1.5/10.8) in terms of an effect size. This upper bound on the African American–specific treatment effect is less than half of the impact estimate reported in the original study. These comparative results imply that the null results in this experiment are precisely estimated and do not concur with the effect sizes reported in the seminal experiments.

Because the processes by which the intervention is thought to operate are "recursive" (e.g., Yeager & Walton, 2011), the effects of the intervention by marking periods was also examined. Half of the participating schools had three marking periods and half had four marking periods. Impact estimates estimated separately by marking periods provide no indication that the treatment led to higher grades or obvious differences in effects across the academic year. Results broken out for each school are not reported here, but they also lead to null findings generally. The only exceptions are that, in one school, there was a small positive effect for African American students ($p = .078$). In this school, there was a positive effect of the affirmation for White students. In a second school, there was a *negative* effect for African American students ($p = .086$) and no statistically significant effect for White students.

Table 4. Determinants of final grade in treated subject

Independent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.090 (0.365)			0.041 (0.370)			
Treatment × White		0.121 (0.631)	−0.011 (0.632)		−0.028 (0.643)	−0.034 (0.643)	0.063 (0.596)
Treatment × Black		0.402 (0.654)	0.276 (0.663)		0.228 (0.646)	0.230 (0.648)	−0.009 (0.599)
Treatment × Hispanic		0.717 (1.060)	0.590 (1.056)		0.582 (1.048)	0.591 (1.049)	0.591 (0.916)
Treatment × Female		−0.492 (0.650)	−0.354 (0.640)		−0.280 (0.629)	−0.279 (0.629)	−0.431 (0.635)
Black	−7.027***(0.671)	−7.166***(0.780)	−7.930***(0.804)	−7.010***(0.641)	−7.136***(0.751)		
Hispanic	−9.525***(0.797)	−9.821***(1.067)	−9.665***(1.053)	−8.835***(0.750)	−9.140***(1.022)		
Female	3.701***(0.451)	3.944***(0.554)	3.920***(0.549)	3.647***(0.450)	3.785***(0.565)		
White female						3.450***(0.713)	2.177**(0.634)
Black female						−3.280***(0.963)	−0.957 (0.808)
Black male						−7.552***(0.850)	−4.626***(0.778)
Hispanic female						−5.623***(1.182)	−2.371*(1.028)
Hispanic male						−9.220***(1.152)	−4.840***(0.8552)
Baseline reading score							5.074***(0.234)
R^2	0.167	0.167	0.192	0.397	0.398	0.398	0.539
School fixed effects	No	No	Yes	No	No	No	No
Classroom fixed effects	No	No	No	Yes	Yes	Yes	Yes

Note. The dependent variable is the final grade in the treated subject. Standard errors are adjusted for heteroscedasticity clustered at the classroom level.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 5 explores multiple possible forms of treatment heterogeneity associated with student traits. In general, these results are similar to those in Table 4 indicating that the treatment had no statistically significant effects across different types of students. For example, the race and ethnicity-specific treatment estimated effects do not vary meaningfully by gender. Furthermore, the treatment appears to have been similarly ineffective for students in the top, middle, and bottom third of the distribution of reading achievement at baseline as well as among students for whom this baseline measure was unavailable. Of interest, the evidence from the original study suggests that the intervention was more effective among students with lower baseline performance.

The bottom panel of Table 5 indicates that the treatment effects were small and statistically insignificant across study years and among seventh graders. However, the final row of Table 5 indicates that, among eighth graders, the treatment led to a statistically significant increase in the performance of Hispanic students (effect size = 0.38) and a statistically significant reduction in the performance of female students (effect size = – 0.21). This pattern of Hispanic and female-specific effects is also apparent in the third of the sample taught by female teachers (i.e., nine of 22 teachers). There is some related evidence of negative effects of the affirmation for female students in the original study by Cohen et al. (2006). Specifically, Cohen et al. (2006, Experiment 1) found that European-American girls responded negatively to intervention, other things being equal ($p < .05$). This negative effect was not found in Experiment 2. However, the full-sample regression estimates based on Experiment 2 and performance in the treated subject (Cohen et al., 2006, Table S2) do indicate a large, negative, and weakly significant interaction between assignment to the affirmation and being female. That is, these results suggest that, other things being equal, the effect of the affirmation on course performance is smaller for females, $d = -0.36$, $t(119) = -1.94$.

Other Outcomes

Table 6 presents the estimated effects of the treatment by race, ethnicity, and gender and conditional on classroom fixed effects on the other available outcome measures. These consistently indicate that the treatment had small and not statistically significant effects on multiple outcomes: grades in other courses, state math and reading assessment scores, absences, tardiness, and disciplinary infractions. Infractions consist of reported student misbehavior (e.g., disruptive or disrespectful behavior, cutting class) regardless of whether it resulted in a formal suspension. Similarly, the last row in Table 6 indicates that the treatment did not have significant effects on the completion of word-fragments in a stereotype-relevant manner. For each of the last four count-data variables, these null findings are replicated in negative-binomial specifications that accommodate classroom fixed effects.

However, models that focus on eighth graders (i.e., not the full-sample results reported in Table 6) generate some findings similar to those in Table 5. For example, for Hispanic students, the treatment led to large increases (effect sizes of about 0.3 to 0.4) in course grades in two other subjects (p of .074 and .011). The treatment also appears to have increased the performance of eighth-grade Hispanic students on the state math assessment by a weakly significant 0.28 standard deviations ($p = .090$). With regard to eighth-grade females, the affirmation appears to have *decreased* final grades in the same two nontreated subjects, which saw gains for Hispanic students (effect sizes in the range of 0.18 to 0.25, p of .039 and .092). However, it should be clearly noted that these statistically significant findings may simply be a spurious reflection of the multiple-comparisons problem (Schochet, 2008).

Table 5. Treatment estimates by student trait

Sample	Treatment Effects by Student Trait				R^2	Sample Size
	White	Black	Hispanic	Female		
Full sample	−0.028 (0.643)	0.228 (0.646)	0.582 (1.048)	−0.280 (0.629)	.398	2348
Girls	0.265 (0.680)	−0.214 (0.831)	−0.185 (1.228)	n/a	.409	1188
Boys	−0.524 (0.761)	0.380 (0.886)	1.143 (1.492)	n/a	.451	1160
Top baseline reading	−0.322 (1.111)	−0.328 (1.337)	−1.462 (1.965)	−0.426 (1.188)	.502	757
Middle baseline reading	−0.299 (1.080)	−0.110 (1.146)	1.122 (1.646)	0.258 (1.208)	.518	741
Bottom baseline reading	0.912 (1.105)	−0.777 (1.157)	2.310 (1.935)	−0.172 (1.346)	.537	703
Missing baseline reading	2.690 (5.536)	7.01 (4.089)	−5.398 (7.831)	−4.870 (5.239)	.664	147
Academic year 2008–09	1.119 (0.730)	0.351 (0.883)	1.695 (1.728)	−1.359 (0.823)	.464	1023
Academic year 2009–10	−0.777 (0.928)	0.387 (0.944)	−0.096 (1.344)	0.544 (0.897)	.352	1325
Grade 7	−0.193 (0.831)	0.154 (0.817)	−1.598 (1.372)	0.891 (0.835)	.354	1463
Grade 8	0.269 (1.040)	0.256 (1.077)	4.093**(1.351)	−2.233*(0.883)	.481	885

Note. The dependent variable is the final grade in the treated subject. All models condition on indicators for race, ethnicity, and sex and on classroom fixed effects. Standard errors are adjusted for heteroscedasticity clustered at the classroom level.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 6. Estimated treatment effects on other outcomes

Dependent Variables	Treatment Effects by Trait				Sample Size
	Treatment × White	Treatment × Black	Treatment × Hispanic	Treatment × Female	
Other Subject A	0.511 (0.634)	0.319 (0.667)	−0.089 (1.046)	−0.511 (0.657)	2, 329
Other Subject B	0.569 (0.658)	0.808 (0.689)	−1.300 (1.209)	−0.615 (0.717)	2, 305
Other Subject C	0.273 (0.637)	0.090 (0.672)	0.505 (1.059)	−0.524 (0.676)	2, 334
Reading score	0.004 (0.057)	−0.039 (0.053)	−0.048 (0.084)	0.042 (0.058)	2, 381
Math score	−0.041 (0.065)	−0.031 (0.060)	−0.046 (0.095)	0.040 (0.060)	2, 376
Absences	0.003 (0.723)	−0.267 (0.808)	−0.211 (0.963)	1.033 (0.758)	1, 799
Tardies	−1.595 (1.123)	0.098 (0.854)	1.017 (0.865)	1.613 (1.024)	1, 799
Disciplinary infractions	−0.087 (0.123)	0.075 (0.209)	0.139 (0.193)	0.003 (0.203)	1, 718
Stereotype-themed words	−0.041 (0.100)	0.070 (0.109)	0.036 (0.134)	0.023 (0.111)	2, 077

Note. All models condition on indicators for race, ethnicity, and sex and on classroom fixed effects. Standard errors are adjusted for heteroscedasticity clustered at the classroom level.

More specifically, Table 6 presents 36 new hypothesis tests based on the full sample. Focusing only on the subset of eighth graders then doubles the number of hypothesis tests. Even when the null of no effect is consistently true, we might expect to find test statistics that reject the null five percent of the time (i.e., (72 × .05) or 3 to 4 times), simply by chance.

Treatment Moderators

The effects of the treatment may be moderated by classroom traits, several of which (e.g., class size, the baseline achievement levels of peers, the racial/ethnic composition of peers) are observable. Such evidence on possibly relevant contextual determinants can be useful not only to engage external-validity concerns but also to inform our understanding of this study's unexpected null findings. The large number of participating classrooms in this study makes it possible to examine the classroom traits that are related to treatment efficacy. To provide such evidence, I estimated—separately for each participating classroom—treatment effects unique to White, African American, Hispanic, and female students. Figure 1 illustrates this variation, showing the 95% confidence for each classroom-level treatment effect for each subgroup ordered from smallest to largest. These classroom-level impact estimates suggest that the intervention's effects, although clearly centered on zero, also exhibited nontrivial variation across classrooms. In particular, for each participating classroom, there are at least a few classrooms where the intervention had statistically significant effects in both positive and negative directions.

To understand this variation in treatment efficacy across classrooms, I estimated linear regressions where these impact estimates were the dependent variables and the key independent variables were observable classroom traits.[6] Table 7 presents the key results of this

[6]To improve the efficiency of these classroom-level regressions, I use the inverse of the standard error for each treatment estimate as a weight. There are fewer than 139 observations in this analysis because some classrooms did not have within-classroom variation with regard to ethnicity, gender, and treatment status.

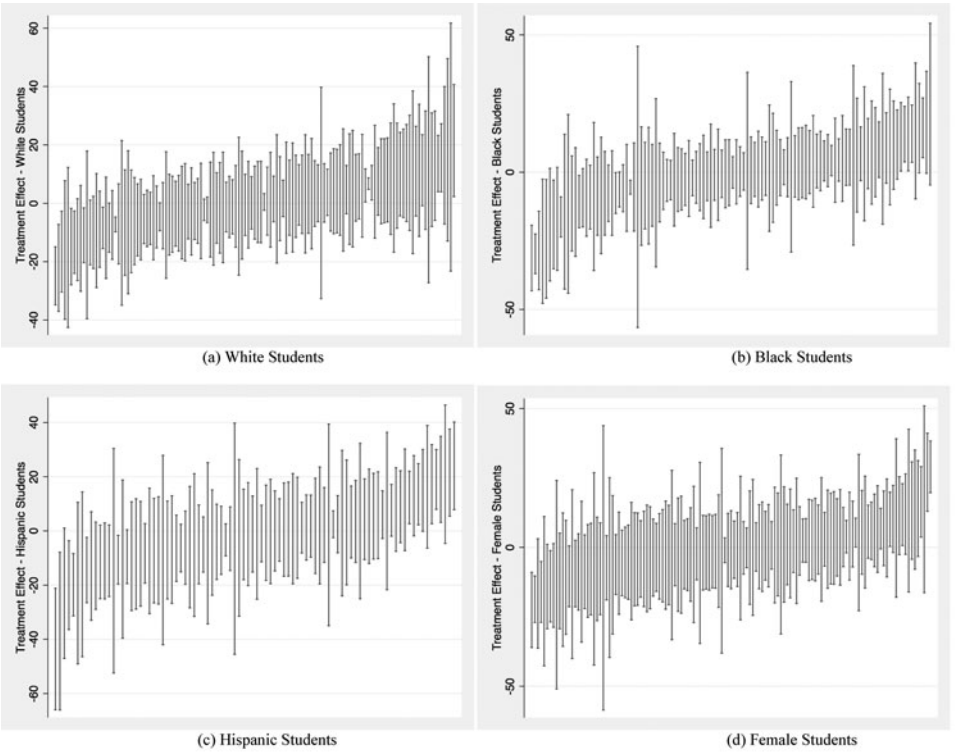

Figure 1. Ninety-five percent confidence interval for classroom-level treatment effects by race, ethnicity, and gender.

Table 7. Predictors of classroom level treatment effects by race, ethnicity, and sex

	By Student Traits			
Independent Variable	White	Black	Hispanic	Female
% Black	2.200 (5.433)	16.486*(7.358)	5.796 (9.609)	−0.006 (5.919)
% Hispanic	−11.012 (8.316)	3.973 (10.499)	17.374 (14.490)	3.278 (8.706)
Class size	0.070 (0.152)	0.200 (0.202)	−0.293 (0.266)	−0.032 (0.165)
Baseline peer achievement	−1.592 (2.337)	−0.509 (2.773)	1.784 (4.144)	1.895 (2.419)
Missing peer achievement	4.585 (9.418)	10.169 (11.188)	2.783 (14.508)	−8.376 (9.478)
Peer achievement growth	1.107 (3.149)	9.233*(4.357)	15.005*(6.414)	−9.582**(3.555)
Treatment take-up rate	−13.951 (8.396)	28.458*(12.544)	−6.145 (15.460)	−21.825*(10.679)
R^2	0.092	0.183	0.105	0.105
Sample size	128	110	91	129

Note. The dependent variables are the estimated White, Black, Hispanic, and female effects of the treatment on the final grade point average in the treated subject, estimated separately for each classroom. These classroom-level models are weighted by the inverse of the standard error for the impact estimate.
$^{*}p < .05$. $^{**}p < .01$.

exercise. In general, most of the observable classroom traits (e.g., class size, baseline peer achievement, percentage Hispanic) were not significantly related to whether the treatment was more effective in that classroom and for all four subgroups. However, there are several notable exceptions. For example, column 2 of Table 7 indicates that the treatment effect for African American students was larger, by a statistically significant margin, in classrooms with larger concentrations of African American students. Specifically, this point estimate implies that, when share of African American students in a classroom increases by 10 percentage points, the effectiveness of the intervention for African American students increased by 1.65 (i.e., roughly 15% of a standard deviation in the outcome measure). An analysis based on the student-level data confirms this heterogeneity. For example, the data from the African American and White students in classes in which 70% or more were African American ($n = 98$) indicate that the affirmation treatment increased the grade performance of African American students ($p = .087$) by an effect size of roughly 0.25.

Table 7 also indicates that, although the *level* of peer achievement did not predict achievement efficacy, the *growth* in peer achievement did. These peer achievement measures are based on the standardized scores from the state reading assessment. These results indicate that, for African American students in classrooms where the peer achievement growth was 0.1 *SD* higher, the intervention was 0.9 more effective (i.e., roughly 0.09 *SD* relative to the outcome measure). The intervention was also significantly more effective for Hispanic students who were in classrooms with stronger growth in peer achievement. This evidence is consistent with the notion that the affirmation is more effective when implemented in environments with "recursive" properties that support and amplify the initial treatment effect (Yeager & Walton, 2011). However, the results in Table 7 also indicate that the treatment had significantly stronger *negative* effects for those girls in classrooms that experienced stronger growth in peer achievement, a pattern I discuss in the next section. The final row of Table 7 examines how the treatment take-up rate predicted the treatment's effects in a given classroom. The take-up rate is defined as the share of students randomly assigned at baseline who completed the first question on the first assigned worksheet. This rate ranges from 66 to 100%. The results in Table 7 indicate that the affirmation had significantly stronger *negative* effects for girls in classrooms where the take-up rate was higher. There is also statistically significant evidence that the affirmation had stronger, positive effects for African American students in classrooms where the take-up rate was higher. The observed take-up rate can reflect both the extent to which a student experienced the treatment contrast and, possibly, the presence of higher performing peers and teachers who might support the impact of the intervention.

DISCUSSION AND CONCLUSIONS

Classroom practices and procedures informed by our growing understanding of the educational consequences of social identity show promise with regard to closing achievement gaps. Furthermore, interventions in this domain have exceptionally low costs and, because of their seeming simplicity, the potential capacity to scale up with relatively high fidelity. However, the field experiment described here largely failed to replicate the earlier findings suggesting that one specific intervention—a simple 15-min writing-based affirmation—could reduce achievement gaps. That is, the main research question in this study generated a null finding: Student completion of the affirmation did not significantly improve course grades, neither overall nor by ethnicity or sex. These findings are unexpected given both the prior evidence of efficacy and the fact the student-directed intervention is a

uniquely simple one whose treatment contrast was, by all indications, implemented with high fidelity and a strong take-up rate. Furthermore, this study had a larger sample size than previous studies and was adequately powered to find educationally important effects.[7]

Nonetheless, there are a number of candidate explanations for this study's null findings. First and most obviously, it may simply be that the affirmation is not generally effective. For example, one possibility is that the character of the treatment contrast weakens the comparative effects of the affirmation. More specifically, the first page of the intervention asks students to identify their most (treatment) or least (control) important values. This prompt implies *both* questions similarly require students to implicitly rank-order their values from the given list. A second possible explanation for this study's null findings turns on external validity concerns. In this study, nearly 90% of eligible students participated. In the study by Cohen et al. (2006), only 50% of students provided consent and participated. If those who provided consent differed systematically in an outcome-relevant way, it could explain why the treatment was apparently effective in the original study.[8]

A third, conjectural explanation that can be more easily dismissed is that the block-randomization procedure exacerbated treatment contamination and attenuated the treatment estimate. That is, it may be that those who were paired with each other prior to randomization (i.e., same race-ethnicity, gender, and similar baseline achievement) are more likely to know each other. The friends and acquaintances whose achievement gains from exposure to the treatment may subsequently increase the outcomes of those who received the control. Although this is theoretically possible, the implied bias would not be empirically relevant even under generous assumptions about the magnitude of peer-group effects. For example, suppose the true effect is 0.34 *SD* as in the study by Cohen et al. (2006). Assume that an exogenous increase in the achievement of a peer increases own achievement by as much as 50%. The implied treatment contamination for control students would then be 0.17 *SD*. The observed, biased impact estimate would be 0.17 *SD*. However, the data from this study indicate that, for African American students, impact estimates this large are not supported by the data.

In combination with the null findings, the ancillary evidence from this experiment suggests two other broad conclusions that might meaningfully inform further research in this important area. One is that the nature of the treatment contrast implied by the affirmation intervention and its underlying theoretical mechanisms may not be well understood. For example, Cohen et al. (2006, Experiment 1) found that, for European American girls, the affirmation significantly reduced performance in the treated course ($d = –0.48$, t statistic $= –2.49$). This study (Table 5) similarly found that the affirmation exercise reduced the performance of female students in eighth grade by a smaller but more precisely estimated amount ($d = –0.21$, t statistic $= –2.53$). Furthermore, the negative effect of the affirmation for the full sample of female students was significantly more pronounced in classrooms where the treatment uptake rate was higher (Table 7).

This evidence that the affirmation may compromise the academic performance of middle-school female students is not easily recognized with the motivating theory of stereotype threat. Instead, these results suggest the possibility that the affirmation and control worksheets may trigger alternative, outcome-relevant mechanisms that are particularly salient for middle-school female students. For example, other studies suggest that

[7]The treatment estimates for African American students in this study were sufficiently precise to reject effects half as large as those reported by Cohen et al. (2006).

[8]However, it should be noted that this study generates similar null findings using the one school-year observation that also used active consent.

uncertainty about social belongingness can have direct implications for the performance of stigmatized students (e.g., Walton & Cohen 2007). It may be that the *control* condition in this study promotes social belongingness by encouraging students to identify and reflect on the personal values of *others,* whereas the affirmation intervention promotes a contrasting emphasis on individualism through an emphasis and reflection on personal values.

A second and particularly important conclusion from this study is to reinforce the arguments made by Yeager and Walton (2011) indicating that the efficacy of social-psychological interventions can depend critically on the presence of supporting contextual circumstances. One element of these relevant contextual circumstances may be the quality of the implementation. Interventions like this affirmation are thought to be effective because they target students' subjective experiences within schools. However, for this to occur, the interventions need to be implemented in a manner that meaningfully engages the relevant psychological mechanisms (and not just as a hollow process). It could be that this study omitted some relevant implementation detail that meaningfully contributed to the efficacy of the original affirmation study. However, it should also be noted the intervention was student directed and relied on a standard teacher script for presenting it. If the implementation still suffered from some undiagnosed problem or unknown underlying mechanisms, the prospects for scaling up this intervention will turn critically on further research more explicitly identifying the active ingredients of the intervention.

Successful social-psychological interventions are also thought to require a complementary and supportive learning environment that can sustain and amplify the effects of the original treatment contrast (Yeager & Walton, 2011). So, another possible explanation for this study's null findings is that there were unobserved, learning-relevant traits (e.g., teacher quality) that were unique to the single school and three teachers in the original study and that supported and amplified the within-classroom treatment contrast created by the student worksheets. This study's finding that the affirmation was significantly more effective for minority students in classrooms that exhibited stronger growth in student achievement (columns 2 and 3 of Table 7) is strongly suggestive of this hypothesis. However, the fact that the treatment had stronger *negative* effects for female participants in these higher performing contexts (final column of Table 7) also implies that the psychological mechanisms triggered by this particular treatment contrast may not be well understood. These results suggest the need for future research to examine both the mechanisms underlying this context as well as the possibly critical, mediating role played by the broader context in which it is situated.

ACKNOWLEDGMENTS

I thank Erica Johnson for excellent research assistance. I also thank Robert L. Jarvis and the Delaware Valley Minority Student Achievement Consortium for their support and advice. I also thank seminar participants at the Federal Reserve Bank of New York, the 2011 CESifo Area Conference on the Economics of Education, and the 2012 SREE and AEFP research conferences.

FUNDING

This material is based upon work supported by the Institute of Education Sciences under grant number #R305A090162 and the Spencer Foundation. Any opinions, findings, and

conclusions or recommendations expressed in this publication are those of the author and do no necessarily reflect the views of the Institute of Education Sciences or the Spencer Foundation.

REFERENCES

Arbuthnot, K. (2009). The effects of stereotype threat on standardized mathematics test performance and cognitive processing. *Harvard Educational Review*, *79*, 448–472.

Aronson, J., & Dee, T. (2011). Stereotype threat in the real world. In T. Schmader & M. Inzlicht (Eds.), *Stereotype threat: Theory, process, and application* (pp. 264–279). Oxford, UK: Oxford University Press.

Aronson, J., & McGlone, M. (2007). Stereotype threat. In T. Nelson (Ed.), *The handbook of prejudice, stereotyping, and discrimination* (pp. 153–178). New York, NY: Guilford.

Bowen, N. K., Wegmann, K. M., & Webber, K. C. (2013). Enhancing a brief writing intervention to combat stereotype threat among middle-school students. *Journal of Educational Psychology*, *105*, 427–435.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, *96*, 988–1012.

Cohen, G., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307–1310.

Cohen, G., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzutoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*, 400–403.

Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology*, *102*, 479–496.

Cook, M. D., & Evans, W. N. (2000). Families or schools? Explaining the convergence in white and black academic performance. *Journal of Labor Economics*, *18*, 729–754.

Cook, P., & Ludwig, J. (1997). Weighing the "burden of 'acting white"': Are there race differences in attitudes towards education? *Journal of Policy Analysis and Management*, *16*, 256–278.

Dee, T. S. (2014). Stereotype threat and the student-athlete. *Economic Inquiry*, *52*, 173–182.

Fordham, S., & Ogbu, J. (1986). Black students' school success: Coping with the burden of 'acting white."' *The Urban Review*, *18*, 176–206.

Fryer, R. G., & Torelli P. (2010). An empirical analysis of "acting white." *Journal of Public Economics*, *94*, 380–396.

Good, C., Aronson, J., & Harder, J. (2008). Problems in the pipeline: Women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, *29*, 17–28.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*, 645–662.

Guryan, J. (2004). Desegregation and black dropout rates. *American Economic Review*, *94*, 919–943.

Hanselman, P., Bruch, S. K., Gamoran, A., & Borman, G. D. (2014). Threat in context school moderation of the impact of social identity threat on racial/ethnic achievement gaps. *Sociology of Education*, *87*, 106–124.

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen, C. J., Blair, S. S., Rouse, D. I., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology*, *106*, 375–389.

Johnson, R. C. (2011). *Long-run impacts of school desegregation and school quality on adult attainments* (Working Paper No. 16664). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w16664

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*, *29*, 371–381.

Kellow, T. J., & Jones, B. D. (2008). The effects of stereotypes on the achievement gap: Reexamining the academic performance of African American high school students. *Journal of Black Psychology*, *34*, 94–120.

Lauer, S., Momsen, J., Offerdahl, E., Kryjevskaia, M., Christensen, W., & Montplaisir, L. (2013). Stereotyped: Investigating gender in introductory science courses. *CBE-Life Sciences Education*, *12*, 30–38.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Miyake, A., Kost-Smith L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*, 1234–1237.

Neal, D., & Johnson, W. R. (1996). The role of pre-market factors in black–white differences. *Journal of Political Economy*, *104*, 869–895.

O'Neill, J. (1990). The role of human capital in earnings differences between black and white men. *Journal of Economic Perspectives*, *4*, 25–46.

Rampey, B. D., Dion, G. S., & Donahue, P. L. (2009). *NAEP 2008 Tends in Academic Progress (*Report *No. NCES 2009–479)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Reber, S. J. (2010). Desegregation and educational attainment for blacks. *Journal of Human Resources*, *45*, 893–914.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*, 336–356.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations (*Report *No. NCEE 2008–4018)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., . . . Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology*, *104*, 591–618.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.

Walton, G. M. & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*, 82–96.

Woolf, K., McManus, I. C., Gill, D., & Dacre J. (2009). The effect of a brief social intervention on the examination of results of UK medical students: A cluster randomized controlled trial. *BMC Medical Education*, *9*(35), 1–15.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*, 267–301.